Project Images and Notes

Sunday, June 10, 2018

1:44 PM

1985 Auto Imports Database Analyses

# Dataset:

<https://www.kaggle.com/ashishpal2702/1985-auto-imports-database-analyses-prediction>

# Objective:

We would like to examine predict the price of a car based on the effect of physical and performance car attributes while holding the insurance risk rating constant. again.

# Variable list sensitive to the objective:

## Physical Car attributes

1. make: [ Manufacturer name eg : alfa-romero, audi, bmw, chevrolet, dodge, honda,isuzu etc. ]
2. num-of-doors: [four, two].
3. body-style: [hardtop, wagon, sedan, hatchback, convertible]
4. engine-location: [front, rear]
5. wheel-base: [continuous from 86.6 120.9]
6. length: [continuous from 141.1 to 208.1]
7. width: [continuous from 60.3 to 72.3]
8. height: [continuous from 47.8 to 59.8]
9. curb-weight: [continuous from 1488 to 4066]

## Performance attributes

1. fuel-type: [diesel, gas]
2. aspiration: [std, turbo]
3. drive-wheels: [4wd, fwd, rwd]
4. engine-type: [dohc, dohcv, l, ohc, ohcf, ohcv, rotor]
5. num-of-cylinders: [eight, five, four, six, three, twelve, two]
6. engine-size: [continuous from 61 to 326]
7. fuel-system: [1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi]
8. bore: [continuous from 2.54 to 3.94]
9. stroke: [continuous from 2.07 to 4.17]
10. compression-ratio: [continuous from 7 to 23]
11. horsepower: [continuous from 48 to 288]
12. peak-rpm: [continuous from 4150 to 6600]
13. city-mpg: [continuous from 13 to 49]
14. highway-mpg: [continuous from 16 to 54]

## 

## Insurance Risk attributes

1. normalized-losses: [average loss payment per insured vehicle year -> continuous from 65 to 256.]
2. symboling: [its assigned insurance risk rating -> [-3, -2, -1, 0, 1, 2, 3]]

## Response Variable

1. Price

## 

## Data Sets

Original

41 data points were missing from Normalized Losses variable. We chose to remove this variable from the dataset in order to preserve as many records as possible. This variable is not reliable enough to keep in the analysis.

/\*Import dataset with formatted columns\*/

data auto;

infile '/home/carollr0/DataSets/Automobile\_data.csv' dlm=',' firstobs=2;

input symboling normalizedlosses

make $ fueltype $ aspiration $ numofdoors $

bodystyle $ drivewheels $ enginelocation $ wheelbase length width height

curbweight enginetype $ numofcylinders $ enginesize fuelsystem $ bore stroke

compressionratio horsepower peakrpm citympg highwaympg price;

run;

Machine generated alternative text:
Alphabetic List of Vuiables and Attributu 
11 
IS 
24 
14 
17 
15 
18 
13 
11 
16 
23 
20 
10 
12 
Height 
width 
Sli. 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
SEST12_ 
I nf.t 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 
s EST32_ 

/\*Descriptive Statistics for Numeric Variables\*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=auto2 chartype mean std min max n nmiss vardef=df;

var symboling wheelbase length width height curbweight

enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg

price;

run;

Machine generated alternative text:
Variable 
symboling 
wheelbase 
length 
Width 
height 
curbweight 
enginesize 
bore 
stroke 
compressionratio 
horsepower 
peakrpm 
citympg 
highwaympg 
price 
Mean 
0 8341463 
98 7565854 
174 0492683 
65 9078049 
53.7248780 
2555 57 
126 9073171 
3.3297512 
3 2554229 
10.1425366 
104 2561576 
5125 37 
25 2195122 
30 7512195 
13207.13 
Std Dev 
1.2453068 
60217757 
12 3372885 
2 1452039 
2 4435220 
520 6802035 
41 6426934 
0.2735387 
0 3167175 
3 9720403 
39 7143688 
479 3345598 
6 5421417 
6.8864431 
7947 07 
Minimum 
-2 0000000 
86 6000000 
141 1000000 
60 3000000 
47 8000000 
1488 00 
61 0000000 
2 5400000 
2 0700000 
7 0000000 
48 0000000 
4150 00 
13 0000000 
16 0000000 
511800 
Maximum 
3.0000000 
120.9000000 
208.1000000 
72 3000000 
59.8000000 
4066 00 
326.0000000 
3.9400000 
4 1700000 
23.0000000 
288.0000000 
6600 00 
49.0000000 
54 0000000 
45400 00 
N 
205 
205 
205 
205 
205 
205 
205 
201 
201 
205 
203 
203 
205 
205 
201 
N Miss 
4 
4 
2 
2 
4 

Cleaned Data - removed all rows/records with missing data

/\*Delete rows with missing data - equates to 10 records removed \*/

data auto\_clean;

set auto;

if nmiss(of \_numeric\_, 1) + cmiss(of \_character\_, '?') then

delete;

run;

/\*Descriptive Statistics for Numeric Variables\*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.AUTO\_CLEAN chartype mean std min max n nmiss vardef=df;

var symboling wheelbase length width height curbweight

enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg

price;

run;

Machine generated alternative text:
Variable 
symboling 
wheelbase 
length 
width 
height 
curbweight 
enginesize 
bore 
stroke 
compressionratio 
horsepower 
peakrpm 
citympg 
highwaympg 
N 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
195 
N Miss 
O 
o 
o 
o 
o 
O 
o 
O 
O 
o 
o 
O 
o 
o 
O 
Mean 
07948718 
98.8964103 
174.2569231 
65.8861538 
538615385 
2559 oo 
1278384615 
3.3293846 
3.2503077 
10.1949744 
103.2717949 
5099.49 
25.3743590 
30.8410256 
13248.02 
std Dev 
1.2306123 
6.1320383 
12.4764434 
2.1324839 
2.3967778 
524.7157994 
41.4339159 
0.2718657 
0.3141145 
4.0621088 
37.8697302 
468.2713809 
6.4013819 
6.8293151 
8056.33 
Minimum 
-2.0000000 
86.6000000 
141.1000000 
60.3000000 
47 8000000 
1488.00 
610000000 
2.5400000 
2.0700000 
7.0000000 
480000000 
4150.00 
13.0000000 
16.0000000 
5118.00 
Maximum 
3.0000000 
120.9000000 
208.1000000 
72.0000000 
59.8000000 
4066.00 
326.0000000 
3.9400000 
4.1700000 
23.0000000 
262.0000000 
6600.00 
49.0000000 
54.0000000 
45400.00 

Deleted Data -

/\*Data set created - Deleted rows with missing data\*/

data auto\_clean\_missing;

set auto\_clean;

if nmiss(of \_numeric\_, 1) + cmiss(of \_character\_, '?') then

output;

run;

./\*Analyze the deleted records\*/

/\*Descriptive Statistics for Numeric Variables\*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=auto\_clean\_missing chartype mean std min max n nmiss vardef=df;

var symboling wheelbase length width height curbweight

enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg

price;

run;

N 
N Miss 
ight 
22222222 
1722832323 
244722 
34300000 
04121182 
54480822 
04400758 
224172" 
815148248 
02614784 
17 zoooooo 
32506410 
41311822 
952000000 
4e zooooŒ 
70000000) 
101 cooooŒ 
13000000) 
22000000) 
55200000) 
132000000) 
135000000) 
22000000) 
31 cocooŒ 

We compared the descriptive stats for Original dataset against the Cleaned dataset and the deleted record dataset.

The means of the deleted items are within range of the original data. We conclude it is safe to delete the 10 records

that contain missing data points.

## Exploratory Analysis

Scatter Plot Matrix of all numerical variables

/\*Plot all numeric variables against each other\*/

options validvarname=any;

ods noproctitle;

ods graphics / imagemap=on;

/\* Scatter plot matrix macro \*/

%macro scatterPlotMatrix(xVars=, title=, groupVar=);

proc sgscatter data=WORK.AUTO\_Clean;

matrix &xVars / %if(&groupVar ne %str()) %then

%do;

group=&groupVar legend=(sortorder=ascending) %end;

diagonal=(histogram normal);

title &title;

run;

title;

%mend scatterPlotMatrix;

%scatterPlotMatrix(xVars=symboling wheelbase length width

height curbweight enginesize bore stroke compressionratio horsepower peakrpm

citympg highwaympg price, title="Clean Dataset - Scatter plot matrix",

groupVar=);

Machine generated alternative text:
Clean Dataset - Scatter plot matrix 
sy whe 
length width height cur 
engL 
bore stroke 
pea_ higtm 
price 
•uaaaaaaooao•sa 
onoaonaaaoao•sa ' 
aanaaasaaaa•ssa ' 
sesssssesgeeess 

proc corr data=auto\_clean plots=matrix (histogram);

run;

Machine generated alternative text:
Scatter Plot Matrix 
symboling 
normalizedlosses 
wheelbase 
length 
width 

Machine generated alternative text:
CO"elation Cæffcients, N = 195 
Prob > n HO: 
hi ympg 
boli ng 
1.00000 
-0.53±7 
-0.51754 
-o .23025 
00012 
-0.06823 
-0.14582 
086-81 
-0.1812B 
0.0112 
0.072" 
02060 
00012 
08704 
-0.08412 
02423 
1.00000 
0.37022 
0.31001 
o.5@250 
0.78272 
0.5070 
0.40823 
0.17172 
0.24773 
00005 
0.37554 
0.5857 g 
-0.36306 
0.37022 
1.00000 
0.35803 
0.88166 
o .0748 
o. 13017 
00253 
0.58381 
-o. dude 
-0.71022 
wi dth 
0.31001 
e 0001 
1.00000 
0.315-83 
0.86721 
e 0001 
0.54421 
e 0001 
0.13643 
000.1 
o. Igloo 
00075 
0.61673 
e 0001 
00004 
-0.64710 
e 0001 
0.75427 
e 0001 
-0.51754 
o.5@250 
0.31583 
1.00000 
e 0001 
06642 
0.18023 
00080 
04407 
00002 
-018441 
02407 
-o .26403 
00002 
-0.10227 
0.13829 
-0.22025 
00012 
0.36721 
0.30773 
0.17279 
0.15523 
0.73029 
-02704 
-0.77217 
-0.31271 
-0.06823 
o.56@70 
o .0748 
o Zd42 
o .3E757 
1.00000 
o.5830@ 
00029 
o .024'2 
03327 
o.8426@ 
-0.21001 
00021 
-0.73214 
-0.14582 
0.40823 
0.54421 
0.18023 
00080 
o.5830@ 
1.00000 
-0.0667 g 
0.00306 
0.5680 
-o. 60004 
0.54687 
086-81 
0.17172 
0.13643 
000.1 
04407 
0.17279 
00029 
-0.0667 g 
0.10004 
-0.06820 
02423 
-0.02764 
05013 
-o.03d45 
y 1024 
-0.1812B 
0.0112 
0.24773 
o. 13017 
00253 
00075 
00002 
0.15523 
03327 
o .00306 
-0.21440 
0.32141 
00002 
0.072" 
0.37554 
0.58381 
0.61673 
02407 
0.73029 
o.8426@ 
0.5680 
0.10004 
-0.21440 
0002B 
1.00000 
0.10±5 
-0.32412 
-0.312.2 
0.22060 
00012 
00004 
00002 
-o. 27 "4 
-0.21001 
00021 
-0.06820 
02423 
-o .44458 
0.10±5 
1.00000 
-0.10423 
c itympg 
08704 
-o. dude 
-0.64710 
-0.77217 
e 0001 
-0.02764 
0.32141 
-0.82412 
1.00000 
-o. 70268 
pg 
-0.71022 
-0.31271 
-0.73214 
-o.03d45 
00002 
-0.81202 
0.97225 
1.00000 
-0.71559 
-0.08412 
02423 
0.58579 
0.75427 
0.13829 
0.33573 
0.388ß4 
0.54687 
y 1024 
0.0050 
-o. 70268 
-0.7159 
1.00000 

Initial look at Correlated variables

Highway MPG and City MPG

Length and Wheelbase

Width and Length

Wheelbase and Cure weight

Width and Curb weight

Engine Size and Curb weight

Engine Size and Horsepower

The following features are being removed because of their high correlation with other features.

Curb Weight

Wheelbase

Highway MPG

## Residual Plots, Outliers and Leverage

Proc univariate data=auto\_clean;

var symboling /\*wheelbase\*/

length width height /\*curbweight\*/

enginesize bore stroke compressionratio horsepower peakrpm

citympg /\*highwaympg\*/;

histogram;

run;

proc print data=work.auto\_clean; run;

|  |  |  |  |
| --- | --- | --- | --- |
| Symboling    Machine generated alternative text: Ext.  -2  -2  -2  172  134    Screen clipping taken: 6/13/2018 8:01 AM | Length    Machine generated alternative text: Ext.  1500  1500  31  3-5  34  Hi ghut  202 z  202 z  2031  48  38  70    Screen clipping taken: 6/13/2018 8:02 AM | Width    Machine generated alternative text: Ext.  44  41  63.4  63. a  Hi ghut  71  71  720  722  37  38  70  71    Screen clipping taken: 6/13/2018 8:02 AM | Height    Machine generated alternative text: Ext.  78  77  Hi ghut  59.1  59.1  59.8  59.8  148  2.  120    Screen clipping taken: 6/13/2018 8:03 AM |
| Enginesize    Machine generated alternative text: Ext.  70  70  70  5-8  57  Hi ghut  304  48  75  74    Screen clipping taken: 6/13/2018 7:59 AM | Bore    Machine generated alternative text: Ext.  2.54  2.91  2.91  2.91  12.  34  2.78  2.30  2.30  70  71  122    Screen clipping taken: 6/13/2018 8:03 AM | Stroke    Machine generated alternative text: Ext.  Hi ghut  2.07  2.1g  2.1g  2.3Z  12.  110  144  31  121  3-0  48  4.17  4.17    Screen clipping taken: 6/13/2018 8:03 AM | Compression ratio    Machine generated alternative text: Ext.  121  114  Hi ghut  22  22  22  22  22  177  17g  137    Screen clipping taken: 6/13/2018 8:04 AM |
| Horsepower    Machine generated alternative text: Ext.  Highut  48  52  52  17g  177  37  154  207  207  207  262  238  122  124  125    Screen clipping taken: 6/13/2018 8:04 AM | Peak RPM    Machine generated alternative text: Ext.  111  107  Hi ghut  doco  doco  dzc.o  3-5  37    Screen clipping taken: 6/13/2018 8:04 AM | City MPG    Machine generated alternative text: Ext.  Highut  3-8  3-8  14  70  48  154  31    Screen clipping taken: 6/13/2018 8:05 AM |  |

/\*Dealing with Many Explanatory Variables

#Check residuals to be normally distributed but not individual variables in matrix,

#Fix any outliers - use proc univariate to point out outliers "Extreme Observations"

\*/

Proc reg data=work.auto\_clean plots(label)=(rstudentbyleverage cooksd);

Model price = symboling length width height enginesize bore stroke compressionratio horsepower peakrpm citympg /VIF;

run;

quit;

REG 
MODELI 
Vuiable: 
of 
Anatysis M Varia. 
11 
183 
104 
e ocai 
C—ted Tota I 
boli ng 
width 
twght 
City mp g 
183-5484708 
12591434187 
13248 
2300557 
1002.72 
Adj R. sq 
var 
0.8542 
0.8455 
pr Itl 
e ocai 
06052 
117.24488 
-42.35121 
130. eoo. 
72277 
-2775.73"3 
2%.1283 
40.43228 
1.%410 
15721 
22t.E0177 
232.17101 
78.84222 
17.0P433 
327 
227 
02873 
00041 
000-82 
e ocai 
00005 
00013 
0.01gc 
0±710 
Infla 
1.50408 
478214 
2.12042 
a. 50852 
2.0P2P1 
1.177" 
8.105a 
1.87428 
314119 

Screen clipping taken: 6/13/2018 8:14 PM

64 
REG 
MODELI 
Vuiable: 
Cook's D for price 
99 
200 
Observation 

Screen clipping taken: 6/13/2018 8:14 PM

Outlier and Leverage Diagnostics for price 
17 14 
c8 
o Oyo 
80003D 0 04 
00 
0 05 
o 
0 
00 
& 120 
107 
015 
Leverage 
0 20 
0 25 
0 30 
Outlier 
o 
Leverage 
o Outlier and Leverage 

Fit Diagnostics for price 
1 5000 
1 0000 
5000 
-5000 
-1 0000 
1 5000 
1 0000 
5000 
-5000 
-1 0000 
14 641% 17 
997* 
20000 40000 
Predicted Value 
1464 121 
107 
03 
67 
50000 
40000 
30000 
20000 
1 0000 
30000 
20000 
1 0000 
-1 0000 
14 121? 07 
0 8 66 
20000 
40000 
Predicted Value 
149 
30 
20000 40000 
Predicted Value 
Fit*vlean Residual 
Proportion Less 
0 
02 
Leverage 
100 ISO 
Obsevwation 
Observations 
Parameters 
Error 
R- Square 
0 8542 
Adj R-square 08455 
200 
-1 0500 
Quantile 
-1 soo 7500 
Residual 

/\*Observations 47, 120 and 121 are shown to be leverage and outliers but only observation 47 had a higher Cooks'D and leverage.Observation 47 will be removed.\*/

data auto\_remobs;

set work.auto\_clean;

if \_N\_=47 then delete;

run;

quit;

proc print data=auto\_remobs; run;

/\*Re-Run residuals with corr

The Cooks'd for observation 47 was .06. The remaining indicated outliers/leverage points are now below 0.1. The remaining outliers will be included as is.

\*/

Proc reg data=work.auto\_remobs plots(label)=(rstudentbyleverage cooksd);

Model price = symboling length width height enginesize bore stroke compressionratio horsepower peakrpm citympg /VIF;

run;

quit;

REG 
MODELI 
V.iable: 
11 
182 
Tota I 
of 
Anatysis 
10424320820 
1340812250 
12071143070 
3008.05ęc.o 
13131 
2220852 
Est 
124 
P0484łg 
Adj R- sq 
104.73 
var 
boli ng 
width 
Esti 
_44.122î1 
350.13430 
257.57313 
144.82743 
-17."0î5 
15121 
218.73823 
221.5łgź5 
132.47752 
13.č4č24 
117004048 
77244180 
75.1510@ 
-Y 54 
10.30 
-157 
Y 82 
315 
ltl 
00005 
o. 7785 
02424 
00025 
o. 0552 
0.1175 
00002 
0001. 
08205 
Infla ti 
7.120E 
4.08224 
2.12101 
o .0428č 
2.10288 
7.44051 
1.3740č 

## 

## 

|  |  |
| --- | --- |
| Cook's D for price  0.10  0.06  004  002  0.00  1 so  100  Observation | Machine generated alternative text: Outlier and Leverage Diagnostics for price  0 00  9  o  0 00  06  800  8  96 0  00  0 05  o  06 s  015  Leverage  67 30  0 20  0 25  0 30  Outlier  o  Leverage  o Outlier and Leverage |

## 

Fit Diagnostics for price 
1 0000 
5000 
-5000 
-1 0000 
1 0000 
5000 
-5000 
-1 0000 
0*ecg0017 
30 
30 
870 
896 
20000 
40000 
Predicted Value 
Quantile 
17m 97 
20000 
001 
40000 
40000 
30000 
20000 
1 0000 
30000 
20000 
1 0000 
-1 0000 
Predicted Value 
14 
30 
20000 
40000 
Predicted Value 
Fit*vlean Residual 
Proportion Less 
119 67 
30 
1699 170 
8 95 
96 94 
Leverage 
ono 
008 
0 06 
004 
0 02 
0 00 
100 150 200 
Observation 
Observations 
Parameters 
Error 
905E6 
R- Square 
0 8636 
Adj R-square 08553 
-1 0000 
Residual 
1 0000 

## 

## 

2n9- コ 25- 、 市 
Va ュ 国b-2 
00 コ も 「 市 5 凸 on 「 国 00 
sy ョ = ng 
sy ョ = ng width 2n9 = - 市 5 2 
も 753 朝 朝 223 朝 , 745B 朝 」 821 
も , D173 朝 」 315 い 2083 も 778 
も 」 43 朝 い 朝 07 心 朝 432 い 2031 
も 日 557 い 朝 28 心 朝 , 73 朝 」 017 
も 朝 355 朝 , 52 い 朝 鬲 552 1 日 0 冒 
ー い 2446 朝 573 1 日 0C0 朝 鬲 552 
と 鬲 朝 C 朝 1 日 0C0 朝 573 朝 , 52 い 
朝 日 053 ー い 朝 85B も 6322 ー 01311 
い 23D2 ー い 28D2 ー い 2524 ー い 2714 
朝 日 21 朝 朝 835 い 朝 043 も 315 
1 日 0 冒 も 3 朝 C 朝 ー い 244 朝 も 5355 
Nu ョ - 、 3- -5 IJæd 124 
Nu ョ - 、 ati 望 -5 Rad 124 
も , 7123 も も み 4 
と 暠 277 ー い 2773 と 7D5 
も C557 も 」 43 朝 も , D173 
朝 PO 叫 朝 478 0 , 112B 
朝 254 朝 , 57 朝 」 414 
い 2 朝 8 朝 も 日 朝 1 朝 1 日 0 冒 
い 朝 015 1 日 0 冒 も 日 朝 1 朝 
朝 」 017 い 2031 も , D778 
朝 73 朝 432 い 2083 
い 朝 28 心 い 朝 07 心 朝 」 315 
1 日 0C0 い 朝 015 い 2 朝 36 
REG P 「 00 市 d に 「 2 
bo 「 市 5 尊 Ok 市 00 ョ を 25 凸 「 国 00 「 52 て 0 市 「 市 ak 「 2 ョ ョ pg 2 ュ 0 市 
ー い 2323 
朝 鬲 3 胃 
い 2038 
朝 日 018 
い 27 冒 
P1587 
1 日 0 冒 
も 83 も , D723 1 日 0 冒 も 朝 255 
と 暠 323 も 転 44 ← 朝 鬲 3E0 0 朝 朝 52 
朝 031 と 」 03 ← と 255 1 日 0 呂 
朝 朝 7g 朝 ー い 2773 と 朝 478 
朝 254 ー い 2277 も , 7123 朝 2D2 
い 朝 043 ー い 2524 も 6322 朝 , 745B 
朝 835 ー い 28D2 ー い 朝 85B 朝 朝 223 
朝 日 2 言 
1 日 0 冒 
い 23D2 朝 日 053 も 753 
1 日 0C0 と 723 と 」 03 ← 
01157 も 3 朝 031 

## 

City MPG and Horsepower show to be highly correlated but the VIF show them to both be 7. We will proceed with both features since horsepower is an important car performance feature.

## 

## Assumptions

Linearity - The relationship between the dependent variable and the continuous independent variables are linear

Normality – Residuals of the linear model are normally distributed

Equal Variance – The variance of the residuals is constant for every combination of independent variables and thus constant across all of the predicted values

Independence – Observations are assumed to be independent based on the study description. The car make is indicated but not the model. All we know is that these are cars from 1985 line. We proceed with caution. In future I would try to obtain the model information for the observation to be certain the observations are independent.

## Finding Subset of Variables

/\*Variable Selection techinques - LARS\*/

Proc glmselect data=work.auto\_remobs seed=12;

Class make numofcylinders;

Model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg | numofcylinders make /

selection=LAR (choose=cv stop=cv) CVDETAILS;

Run;

quit;

Fit Statistics

The optimal model using LARS variable selection has 9 features selected. Although the Adj R Square shows that 85% of the variation in price is explained by the 9 features. We will look at LASSO variable selection model next.

GLMSELECT Pr—fure 
V. ble 
Stop 
Cross 
Cross Fold 
Effæt Enforæd 
Numb« SQd 
of Rad 
of Llæd 
Number of 
of 
GLMSELECT Pr—fure 
LAR Sny 
Ste p 
Ente red 
height 
• of 
REM08S 
CV PRESS 
24±740173 
2314178447 
2043847207 
10 
stogg.d st minimum 'f PRESS. 
Stop Details 
nd idate 
CV PRESS 
CV PRESS 

Machine generated alternative text:



pape•es 
JOJ puenbs u0!ssu60Jd 

GLMSELECT Pr—fure 
Widatim. is at Step S 
InteQgt width *3krpm 
of 
184 
Co Tota I 
Sum of 
10400347107 
12071143070 
128.06 
0031500 
MSE 
C.endent 
Adj R-sq 
CV PRESS 
Cross 
13131 
3314.33203 
atio 
F Left CWt 
42 
38 
37 
Pmte r E 
Inte ræpt 
p rabo 
CV PRESS 
263687575 
320±3730 
18400385B 
-50175 
527.875727 
0.053764 
-1428.617238 
253.504438 

/\*Variable Selection techinques - LASSO\*/

proc glmselect data=work.auto\_remobs plots(stepaxis=number)=(criterionpanel ASEPlot) seed=1;

partition fraction(test=.5);

Model price=symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=lasso(choose=cv stop=cv) CVDETAILS;

run;

Fit Statistics

The optimal model using LASSO variable selection has 4 features selected. The Adj R Square shows that 73% of the variation in price is explained by the 4 features. Although the Adj R Square is lower than the LARS model, the SBC is smaller and the interpretation simpler with 4 features.

GLMSELECT 
Va riable 
Stop C rite 
Cross 
Cross Fold 
Effect Enforced 
Number Sæd 
of 
of U æd 
of IJæd for 
of IJæd for 
of Effects 
of Pa—teÆ 
REvoss 
VALidEti 
100 
15 

GLMSELECT 
LASSO S. ry 
Ste p 
Ente red 
Interæpt 
Candidate 
Ned 
Effects In 
ASE 
4285508B z 
3007370. 
142243431 
107379131 
Test ASE 
8430.1502 
26707181 
260213052 
13464911>0 
CV PRESS 
434305501 
123387230 
• Valæ of 
st minimum of PRESS 
Stop Details 
C date 
CV PRESS 
12244072 
CV PRESS 
1223.3306 

Machine generated alternative text:



Progression of Average Squared Errors by Role for price 
Selected Step 
80000000 
60000000 
40000000 
20000000 
Step 
Training 

SS3Ed > 0 1n0 
E D282 'Z 
ZEEdZtZCt8 
ELdZt DELL 
03Ea2 22 
0 Luns 
a Oda 
(u 一 E 」 匕 SY 
SS3Ed > 0 
ASR 100H 

/\*Run Selected Model (LASSO)\*/

Proc reg data=work.auto\_remobs;

Model price= width curbweight enginesize horsepower /partial;

run;

quit;

Assumptions

Leverage is at .30 which is good.

QQ Plot shows normally distributed residuals.

Partial plots show curbweight to not have a significant slope.

The p value for curbweight shows to be .39. We will remove this feature and re-run proc reg.

Root MSE 
Depende'tV•n•bIe: price 
REG 
Mode' ODEL 一 
一 3 一 'AdjR-Sq a32 
一 385 803 

scoo 
-3 
20000 40000 
Predicted Value 
Fit Diagnostics for price 
Predicted Value 
40000 
predicted Value 
an Residual 
Proportion Less 
o 
000 
020 
ons 
0.10 
oos 110 01S 
Leverage 
SO ISO 
Observation 
194 
0.20 
200 
Quantile 
-ISO 6250 137SO 
Residual 
Parameters 
Error OF 
1.11E7 
R. square 0.8263 
Adj *Square 08227 
-8750 

P 刨 Dependent Resldual 
•0 05 0 00 0 05 
Intercept 
• IC 00 
• IC 00 
IC 00 
15000 
IC 00 
P34 Plots for price 
Pa 彐 3 - Regressor Residual 
, 5Q0 0 5Q0 10Q0 
curbweight 

/\*Final model selection\*/

Proc reg data=work.auto\_remobs;

Model price= width enginesize horsepower /partial;

run;

quit;

Assumptions

Leverage is at .30 which is good.

QQ Plot shows normally distributed residuals.

REG Pr—fure 
Model: MODELI 
Variable: p. 
Number of 
s Rud 
Number of 
of 
Sum of 
12071143070 
Co Tota I 
Pr>ltl 
MSE 
C.endent 
C æff Var 
DF 
Estimate 
738 N2d4 
Adj R-sq 
dud 
ETO r 
1288050 
o. 8257 
0.8220 
-334 
438 
404 

Fit Diagnostics for price 
o 0 0 00 
8 00 
-3 
20000 40Üj 
Predicted Value 
Quantile 
Residual 
0 00 
Predicted Value 
20000 
predicted Value 
an Residual 
Proportion Less 
oos 010 
Leverage 
SO 100 ISO 
Observation 
194 
200 
Parameters 
Error OF 
1.11E7 
R. square 0.8257 
Adj *Square 08229 

